



Docket No. AM999074

In re application of: Anita Wai-Ling Huang et al.
Serial No.: 09/513,058
Filed: February 24, 2000
For: SYSTEM AND METHOD FOR CLASSIFYING
ELECTRICALLY POSTED DOCUMENTS

COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

Transmitted herewith, in triplicate, is Appellants' Brief in support of their appeal to the Board of Patent Appeals and Interferences from the Examiner's final rejection in the Office Action dated July 26, 2005.

☐ A petition for extension of time is enclosed.

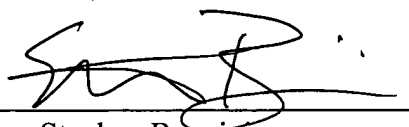
☒ The Commissioner is hereby authorized to charge payment in the amount of \$ 500.00 to cover the filing fee to Deposit Account No. 09-0441.

☐ The Commissioner is hereby authorized to charge payment in the amount of \$ _____ to cover the extension fee to Deposit Account No. 09-0441.

☒ The Commissioner is hereby authorized to charge payment of any necessary fees associated with this communication or credit any overpayment to Deposit Account No. 09-0441.

Respectfully submitted,

Date: January 4, 2006


Stephen Bongini

Registration No. 40,917

FLEIT, KAIN, GIBBONS,
GUTMAN, BONGINI & BIANCO P.L.
One Boca Commerce Center
551 Northwest 77th Street, Suite 111
Boca Raton, Florida 33487
Telephone: (561) 989-9811
Facsimile: (561) 989-9812



AF JFW
PATENT

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:)
Anita Wai-Ling Huang et al.)
Serial No.: 09/513,058)
Filed: February 24, 2000)
Examiner: A. Basehoar)
Group Art Unit: 2178)
For: SYSTEM AND METHOD FOR)
CLASSIFYING ELECTRICALLY)
POSTED DOCUMENTS)
_____)

APPELLANTS' BRIEF

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

Appellants hereby respectfully submit this brief in support of their appeal to the Board of Patent Appeals and Interferences from the Examiner's final rejection in the Office Action dated July 26, 2005.

CERTIFICATE OF MAILING

I hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail in an envelope addressed to: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450

on: 1/4/06 by: Stephen Bongini
DATE

01/24/2006 LWONDIM1 00000016 090441 09513058

01 FC:1402 500.00 DA

AM999074

1

09/513,058

I. REAL PARTY IN INTEREST

The real party in interest is International Business Machines Corporation (IBM) of Armonk, NY.

II. RELATED APPEALS AND INTERFERENCES

There are no related appeals or interferences.

III. STATUS OF CLAIMS

Claims 1, 2, 5-12, 14, 17, 18, 20, 21, 25, and 26 are pending. Claims 1, 2, 5-12, 14, 17, 18, 20, 21, 25, and 26 were finally rejected in the Office Action dated July 26, 2005, and are on appeal.

Attached hereto is an Appendix containing a copy of claims 1, 2, 5-12, 14, 17, 18, 20, 21, 25, and 26, which are the claims involved in this appeal.

IV. STATUS OF AMENDMENTS

Appellants have not filed any amendments subsequent to the final rejection in the Office Action dated July 26, 2005.

V. SUMMARY OF CLAIMED SUBJECT MATTER

One preferred embodiment of the present invention includes a metadata parser module 165 coupled to receive electronically posted documents, as shown in FIG. 1B. Specification at page 4, lines 24-26. The metadata parser 165 outputs a metadata summary 200 for each of the posted documents, as shown in FIGs. 2A and 2B. Specification at page 4, lines 24-26; page 5, lines 20-28; page 7, lines 8-9. Each of the metadata summaries has multiple sub-trees 234 and 236. Specification at page 5, lines 29-30. Further, each of these sub-trees 234 and 236 includes multiple nodes. Specification at page 6, lines 3-6; page 6, line 27 to page 7, line 4.

The preferred embodiment also includes a summary repository 170 coupled to receive and store the metadata summaries 200, and a summary consolidator 180 that is coupled to the summary repository 170. Specification at

page 4, lines 28-30; page 5, lines 8-9. The summary consolidator 180 compares a first of the metadata summaries 200 and a second of the metadata summaries on a structural level by comparing the structure of the sub-trees 234 and reference 236 of the first metadata summary with the structure of the sub-trees of the second metadata summary. Specification at page 5, lines 8-9; page 8, lines 5-10. The summary consolidator 180 identifies the first and second documents corresponding to the first and second metadata summaries 200 as distinct if the structures of the sub-trees 234 and reference 236 of the first and second metadata summaries are not equivalent. Specification at page 8, lines 11-18. If the structures of the sub-trees 234 and 236 of the first and second metadata summaries 200 are equivalent, the first and second metadata summaries are furthered compared. Specification at page 8, lines 19-20.

In particular, the first and second metadata summaries 200 are further compared on a textual level by comparing textual content from a first document that is contained in the sub-trees 234 and 236 of the first metadata summary 200 with textual content from a second document that is contained in the sub-trees of the second metadata summary. Specification at page 8 line 28 to page 9, line 3. The first and second documents corresponding to the first and second metadata summaries 200 are identified as distinct if the textual content within the sub-trees 234 and 236 of the first and second metadata summaries are not equivalent. Specification at page 8 lines 11-13; page 9, lines 7-8.

Additionally, a first subset of the metadata summaries is grouped into a first summary group. Specification at page 7, lines 13-14. The first summary group consists of all of the metadata summaries having a first mime-type designation. Specification at page 7, lines 14-15. Further, a first metadata summary and a second metadata summary are selected from the first summary group. Specification at page 8, lines 1-2.

VI. GROUNDS OF REJECTION TO BE REVIEWED ON APPEAL

A. The rejection of claims 1, 2, 5-9, 12, 14, 17, 18, 20 and 21 under 35 U.S.C. §103(a) as being unpatentable over *Brown et al.* (U. S. Patent No. 5,913,208).

B. The rejection of claims 10, 11, 25 and 26 under 35 U.S.C. §103(a) as being unpatentable over *Brown et al.* in view of *Microsoft Press Computer Dictionary* (Microsoft Press, 1997, p. 309).

VII. ARGUMENT

A. CLAIMS 1, 2, 5-9, 12, 14, 17, 18, 20 AND 21 ARE PATENTABLE
OVER *BROWN ET AL.*

The Examiner rejected claims 1-2, 5-9, 12, 14, 17-18, 20-21 under 35 U.S.C. § 103(a) as being unpatentable over *Brown et al.* Appellants respectfully submit that claims 1-2, 5-9, 12, 14, 17-18, 20-21 are patentable over *Brown* because the *Brown* reference does not teach or suggest the claimed limitations of generating a first metadata summary for said first document and a second metadata summary for the second document, wherein the first metadata summary includes a first plurality of sub-trees and the second metadata summary includes a second plurality of sub-trees, and each of the sub-trees includes a plurality of nodes.

Brown also fails to teach or suggest the claimed limitations of comparing the first and second metadata summaries on a structural level by comparing a structure of the sub-trees of the first metadata summary with a structure of the sub-trees of the second metadata summary, and performing a further comparison of the first and second metadata summaries if the structures of the sub-trees of the first and second metadata summaries are equivalent, wherein if the structures of the sub-trees of the first and second metadata summaries are equivalent, performing a further comparison of the first and second metadata summaries that includes comparing the first and second metadata summaries on

a textual level by comparing textual content from the first document that is contained in the sub-trees of the first metadata summary with textual content from the second document that is contained in the sub-trees of the second metadata summary.

Independent claim 1 is representative of independent claims 1, 12, and 17. Independent claim 1 is directed to a method for classifying electronically posted documents that includes:

receiving a first document and a second document;
generating a first metadata summary for said first document and a second metadata summary for the second document, wherein the first metadata summary includes a first plurality of sub-trees and the second metadata summary includes a second plurality of sub-trees, and wherein each of the sub-trees includes a plurality of nodes;

comparing the first and second metadata summaries on a structural level by comparing a structure of the sub-trees of the first metadata summary with a structure of the sub-trees of the second metadata summary;

identifying the first and second documents as distinct if the structures of the sub-trees of the first and second metadata summaries are not equivalent;

if the structures of the sub-trees of the first and second metadata summaries are equivalent, performing a further comparison of the first and second metadata summaries,

wherein the further comparison of the first and second metadata summaries includes the sub-steps of:

comparing the first and second metadata summaries on a textual level by comparing textual content from the first document that is contained in the sub-trees of the first metadata summary with textual content from the second document that is contained in the sub-trees of the second metadata summary; and

identifying the first and second documents as distinct if the textual content within the sub-trees of the first and second metadata summaries are not equivalent.

The *Brown* reference is directed to a document collection of one or more documents and one or more indexes that each include an inverted file with one or more terms. Each of the terms is associated with one or more document

identifiers. The index further includes a document catalog that associates each of the document identifiers with one or more attributes, either intrinsic or non-intrinsic. A search engine process produces a hit list having one or more hit list entries. Each hit list entry, with one or more hit list attributes, is associated with one of the documents that is determined by the search engine to be relevant to the query. A formatter selects one or more of the hit list attributes identified by a hit list attribute selector, and then compares the selected attributes of two or more entries on the hit list to determine whether or not documents associated with these entries are duplicate instances of one another. The determination can be made without examining the content of the document associated with the entries.

With regard to the second limitation of claim 1, Appellants traverse the Examiner's position that the *Brown* reference discloses generating a metadata summary for a first and second document, wherein each of the metadata summaries includes a plurality of sub-trees that comprise a plurality of nodes. The Examiner compares *Brown's* intrinsic and non-intrinsic attributes with the recited sub-trees, and also compares the recited nodes with *Brown's* intrinsic attributes (relevance, title, size) and non-intrinsic attributes (location which includes filename). The cited portions of the *Brown* reference are limited to a document catalogue 250 (FIG. 2B) and a hit-list 350 (FIG. 3B). The *Brown* reference teaches that the document catalogue 250 "contains an entry 290 for every document (see 140, FIG. 1) indexed". See *Brown* at col. 6, lines 35-36. A document catalogue entry stores attributes, which are either intrinsic or non-intrinsic, of the corresponding document. See *Brown* at col. 6, lines 36-41.

The *Brown* reference further teaches that intrinsic attributes 275 "are properties of the document that are established at the time the document is created and that are invariant with a location and replication of the document...such as title 280 and size 285". See *Brown* at col. 6, lines 42-46. Intrinsic attributes can also be a score "that is a function of one or more other

intrinsic attributes". These other intrinsic attributes, which are typically on the hit-list, can include "document length, title, concepts, author, date of publication, and abstract". See *Brown* at col. 6, lines 47-60. The *Brown* reference teaches that non-intrinsic attributes 265 "are properties of the document that vary with respect to one or more document instance...such as location". See *Brown* at col. 7, lines 7-16. The location attribute 220 of the hit-list 350 is taken from the document catalogue 250. See *Brown* at col. 7, lines 53-54. The *Brown* reference also teaches that the set of hit-list attributes 360, 370, and 380 are distinct from the sets of attributes 260, 265, and 275. See *Brown* at col. 7, lines 50-52.

The document catalogue 250 and hit-list 350 and included intrinsic and non-intrinsic attributes referred to by the Examiner are not a metadata summary as recited in claim 1 (at least in accordance with the meaning given to that term in the present invention, see, for example, FIGs. 2A and 2B). Nowhere does the *Brown* reference teach or suggest "generating a first metadata summary for a first document and a second metadata summary for a second document, wherein the first metadata summary includes a first plurality of sub-trees and the second metadata summary includes a second plurality of sub-trees, and wherein each of the sub-trees includes a plurality of nodes". The *Brown* reference only teaches that an attribute table is made with each row containing the attribute vales for one document, as shown in Figure 3B.

With regard to the remaining limitations of claim 1, Appellants respectfully traverse the Examiner's position that *Brown* discloses:

- comparing the first and second metadata summaries on a structural level by comparing a structure of the sub-trees of the first metadata summary with a structure of the sub-trees of the second metadata summary;

- identifying the first and second documents as distinct if the structures of the sub-trees of the first and second metadata summaries are not equivalent;

- if the structures of the sub-trees of the first and second metadata summaries are equivalent, performing a further comparison of the first and second metadata summaries,

wherein the further comparison of the first and second metadata summaries includes the sub-steps of:

- comparing the first and second metadata summaries on a textual level by comparing textual content from the first document that is contained in the sub-trees of the first metadata summary with textual content from the second document that is contained in the sub-trees of the second metadata summary; and

- identifying the first and second documents as distinct if the textual content within the sub-trees of the first and second metadata summaries are not equivalent.

The portions of the *Brown* reference cited by the Examiner disclose a system in which only attribute values are compared to determine if two documents are equivalent. More specifically, in the system of *Brown*, an attribute table is made with each row containing the attribute values for one document, as shown in Figure 3B. Two documents are determined to be identical if certain predefined attributes are the same for the two documents. If the predefined attributes differ, the documents are determined to be distinct.

In contrast, in embodiments of the present invention, a metadata summary is generated for each document, with the metadata summary including multiple sub-trees, and each of the sub-trees including multiple nodes. To determine if two documents are equivalent, their metadata summaries are compared on a structural level by comparing the structures of the sub-trees of the document's metadata summaries. If the structures of the sub-trees of the metadata summaries are not equivalent, the two documents are identified as distinct. As discussed above, the *Brown* reference does not teach generating, for each document, a metadata summary that includes sub-trees with nodes. Thus, the *Brown* reference cannot possibly teach comparing the structure of sub-trees generated for two documents to determine if the documents are distinct. Because embodiments of the present invention first compare the metadata summaries on a structural level, documents with different structures can be quickly identified as distinct without the need to compare attribute values or

textual content. In the system of *Brown*, document attributes must always be compared.

Further, as recited in claim 1, if the structures of the sub-trees are equivalent, the metadata summaries are compared on a textual level by comparing textual content from the documents that is contained in the sub-trees of the metadata summaries. The two documents are identified as distinct if the textual content from the documents that is contained in the sub-trees of the metadata summaries are not equivalent. The *Brown* reference does not teach comparing textual content from documents to determine if the documents are distinct. The *Brown* reference only compares attribute values in determining document equivalency and not textual content from the documents. In fact, the *Brown* reference teaches away from comparing any actual textual content of the documents. See, for example, *Brown* at Title, Abstract, and generally.

Furthermore, the Examiner has taken the position that the recited limitation of “comparing the first and second metadata summaries on a textual level” is disclosed by *Brown* because *Brown* teaches “comparing relevance score (Fig. 3B: 375) which was calculated by information retrieval algorithm as a function of the query (Fig. 3A) and the contents of the document (column 7, lines 55-67)”. However, the relevance score 375 is not textual content from the documents. Claim 1 recites that the metadata summaries are compared on a textual level by comparing textual content from the documents. Because the metadata summaries are compared on a textual level if the structures of the sub-trees are equivalent, documents with the same structure and attributes but with different textual content are correctly identified as distinct. In the system of *Brown*, documents with the same attributes but with different textual content are incorrectly identified as the same.

Additionally, as recognized by the Examiner, *Brown* does not teach comparing the metadata summaries on a structural level before comparing the

textual content. However, Appellants respectfully traverse the Examiner's position that "it would have been obvious to one of ordinary skill in the art at the time of the invention for Brown et al to have compared the sub-tree structures of the metadata summaries of the documents to determine document distinctness before comparing the textual content of the documents to determine distinctness, because by checking the structure of sub-trees first, it could more quickly be established with minimum amount of comparison that the compared documents were distinct from one another". The *Brown* reference does not teach or suggest comparing textual content from the documents and therefore, this claim element is not rendered obvious by the *Brown* reference.

Appellants respectfully submit that independent claims 12 and 17 include similar limitations to those described above with respect to independent claim 1. Accordingly, independent claims 12 and 17 distinguish over *Brown* for at least the same reasons. Furthermore, claims 2, 5-9, 14, 18, and 20-21 depend from claims 1, 12, and 17. Because these dependent claims contain all the limitations of their respective independent claims, claims 2, 5-9, 14, 18, and 20-21 distinguish over *Brown* as well. Therefore, it is respectfully submitted that the rejection of claims 1-2, 5-9, 12, 14, 17-18, and 20-21 should be reversed.

B. CLAIMS 10, 11, 25 and 26 ARE PATENTABLE OVER *BROWN ET AL.* IN VIEW OF *MICROSOFT PRESS COMPUTER DICTIONARY*

The Examiner's rejected claims 10, 11, 25 and 26 under 35 U.S.C. § 103(a) as being unpatentable over *Brown et al.* in view of Microsoft Press Computer Dictionary ("*Microsoft*"). Appellants respectfully submit that claims 10, 11, 25 and 26 are patentable over *Brown* in combination with *Microsoft*. As explained above with respect to independent claim 1, the *Brown* reference does not teach or suggest the claimed limitations of generating a metadata summary for each of the received documents, "herein each of the metadata summaries includes a plurality of sub-trees and each of the sub-trees includes a plurality of

nodes. The arguments made above regarding these similar claim elements are likewise applicable here and will not be repeated for the sake of brevity.

Further, the *Brown* reference, alone and/or in combination with *Microsoft*, does not teach or suggest the claimed limitations of grouping a first subset of the metadata summaries into a first summary group, wherein the first summary group consists of all of the metadata summaries having a first mime-type designation. Additionally, the *Brown* reference, alone and/or in combination with *Microsoft*, does not teach or suggest selecting a first metadata summary and a second metadata summary from the first summary group, or comparing the first and second metadata summaries on a structural level by comparing a structure of the sub-trees of the first metadata summary with a structure of the sub-trees of the second metadata summary, or identifying the first and second documents as distinct if the structures of the sub-trees of the first and second metadata summaries are not equivalent.

Independent claim 10 is directed to a method for classifying electronically posted documents that includes:

grouping a first subset of the metadata summaries into a first summary group, the first summary group consisting of all of the metadata summaries having a first mime-type designation;
selecting a first metadata summary and a second metadata summary from the first summary group;

With regard to the second, fifth, and sixth limitations of claim 10, Appellants traverse the Examiner's position that the *Brown* reference discloses "generating a metadata summary for each of the received documents, wherein each of the metadata summaries includes a plurality of sub-trees and each of the sub-trees includes a plurality of nodes" and "comparing the first and second metadata summaries on a structural level by comparing a structure of the sub-trees of the first metadata summary with a structure of the sub-trees of the

second metadata summary; and identifying the first and second documents as distinct if the structures of the sub-trees of the first and second metadata summaries are not equivalent". The arguments made above with respect to the similar claim limitations of independent claim 1 are likewise applicable here and will not be repeated for the sake of brevity.

With regard to the third and fourth limitations of claim 10, Appellants traverse the Examiner's position that the *Brown* reference in combination with *Microsoft* discloses "grouping a first subset of the metadata summaries into a first summary group, the first summary group consisting of all of the metadata summaries having a first mime-type designation; selecting a first metadata summary and a second metadata summary from the first summary group".

As recognized by the Examiner, the *Brown* reference "does not teach that when receiving a plurality of documents to group the metadata hit-list summaries according to the mime-type designation of each document". However, the Appellants respectfully traverse that Examiner's position that "it would have been obvious to one of ordinary skill in the art, to have used Brown et al method for identifying duplicate documents from search results without comparing document content and grouping the hit-list summaries by MIME-type, because documents of different MIME-types wouldn't have to be compared as their intrinsic attributes would be wholly different, i.e. a regular text/plain MIME-type couldn't be a duplicate of a text/html MIME-type, and thus this would increase efficiency by significantly reducing the number of hit-list summary comparisons".

Nowhere does *Brown* teach or suggest grouping a subset of the metadata summaries (e.g., a first metadata summary and a second metadata summary) into a first summary group and then selecting, for comparison, a first metadata summary and a second metadata summary out of the grouped subset of metadata summaries. *Brown* is completely silent with respect to what is recited in this claim limitation. Therefore, the *Brown* reference does not teach or suggest

"grouping a first subset of the metadata summaries into a first summary group, the first summary group consisting of all of the metadata summaries having a first mime-type designation; selecting a first metadata summary and a second metadata summary from the first summary group".


Independent claim 25 includes similar limitations as those described above with respect to independent claim 10. Accordingly, independent claim 25 also distinguishes over *Brown*, alone or in combination with *Microsoft*, for at least the same reasons. Claims 11 and 26 depend from claims 10 and 25, respectively. Because these dependent claims contain all the limitations of the independent claims, claims 11 and 26 also distinguish over *Brown*. Therefore, it is respectfully submitted that the rejection of claims 10, 11, 25 and 26 should be reversed.

In view of the foregoing, it is respectfully submitted that the application and the claims are in condition for allowance. Reversal of the final rejection of claims 1, 2, 5-12, 14, 17, 18, 20, 21, 25, and 26 is respectfully requested.

Respectfully submitted,

Dated: January 4, 2006

By: _____


Stephen Bongini
Registration No. 40,917
Attorney for Appellants

FLEIT, KAIN, GIBBONS,
GUTMAN, BONGINI & BIANCO P.L.
One Boca Commerce Center
551 Northwest 77th Street, Suite 111
Boca Raton, Florida 33487
Telephone: (561) 989-9811
Facsimile: (561) 989-9812

VIII. CLAIMS APPENDIX

1. A method for classifying electronically posted documents, the method comprising:

receiving a first document and a second document;

generating a first metadata summary for said first document and a second metadata summary for the second document, wherein the first metadata summary includes a first plurality of sub-trees and the second metadata summary includes a second plurality of sub-trees, and wherein each of the sub-trees includes a plurality of nodes;

comparing the first and second metadata summaries on a structural level by comparing a structure of the sub-trees of the first metadata summary with a structure of the sub-trees of the second metadata summary;

identifying the first and second documents as distinct if the structures of the sub-trees of the first and second metadata summaries are not equivalent;

if the structures of the sub-trees of the first and second metadata summaries are equivalent, performing a further comparison of the first and second metadata summaries,

wherein the further comparison of the first and second metadata summaries includes the sub-steps of:

comparing the first and second metadata summaries on a textual level by comparing textual content from the first document that is contained in the sub-trees of the first metadata summary with textual content from the second document that is contained in the sub-trees of the second metadata summary; and

identifying the first and second documents as distinct if the textual content within the sub-trees of the first and second metadata summaries are not equivalent.

2. The method of claim 1, wherein the further comparison of the first and second metadata summaries further includes the sub-steps of:

before comparing the first and second metadata summaries on a textual level, comparing the first and second metadata summaries on an attribute level by comparing attribute values within the sub-trees of the first metadata summary with attribute values within the sub-trees of the second metadata summary; and

identifying the first and second documents as distinct if the attribute values within the sub-trees of the first and second metadata summaries are not equivalent.

5. The method of claim 1, further comprising identifying the first and second documents as duplicates if the textual content within the sub-trees of the first and second metadata summaries are equivalent.

6. The method of claim 5, further comprising removing the second metadata summary if the first and second documents are identified as duplicates.

7. The method of claim 1, further comprising:

defining a first equivalence metadata table comprising:

a first row corresponding to the first metadata summary;

a second row corresponding to the second metadata summary;

a first column corresponding to the first metadata summary; and

a second column corresponding to the second metadata summary,

wherein the step of identifying the first and second documents as distinct if the structures of the sub-trees of the first and second metadata summaries are not equivalent comprises storing a zero value in the first row and second column position of the first equivalence metadata table.

8. The method of claim 2, further comprising:
defining a first equivalence metadata table comprising:
a first row corresponding to the first metadata summary;
a second row corresponding to the second metadata summary;
a first column corresponding to the first metadata summary; and
a second column corresponding to the second metadata summary,
wherein the step of identifying the first and second documents as distinct if the attribute values within the sub-trees of the first and second metadata summaries are not equivalent comprises storing a zero value in the first row and second column position of the first equivalence metadata table.
9. The method of claim 7, wherein the step of identifying the first and second documents as distinct if the textual content within the sub-trees of the first and second metadata summaries are not equivalent comprises storing a zero value in the first row and second column position of the first equivalence metadata table.

10. A method for classifying electronically posted documents, the method comprising:

receiving a plurality of documents;

generating a metadata summary for each of the received documents, wherein each of the metadata summaries includes a plurality of sub-trees and each of the sub-trees includes a plurality of nodes;

grouping a first subset of the metadata summaries into a first summary group, the first summary group consisting of all of the metadata summaries having a first mime-type designation;

selecting a first metadata summary and a second metadata summary from the first summary group

comparing the first and second metadata summaries on a structural level by comparing a structure of the sub-trees of the first metadata summary with a structure of the sub-trees of the second metadata summary; and

identifying the first and second documents as distinct if the structures of the sub-trees of the first and second metadata summaries are not equivalent.

11. The method of claim 10, wherein the step of grouping further comprises grouping a second subset of the metadata summaries into a second summary group, the second summary group consisting of all of the metadata summaries having a second mime-type designation.

12. A system for classifying electronically posted documents, the system comprising:

- a metadata parser module coupled to receive electronically posted documents, the metadata parser configured to output a metadata summary for each of the posted documents, wherein each of the metadata summaries comprises a plurality of sub-trees and each of the sub-trees includes a plurality of nodes;

- a summary repository coupled to receive and store the metadata summaries; and

- a summary consolidator coupled to the summary repository, the summary consolidator configured to:

 - compare a first of the metadata summaries and a second of the metadata summaries on a structural level by comparing a structure of the sub-trees of the first metadata summary with a structure of the sub-trees of the second metadata summary;

 - identify the first and second documents corresponding to the first and second metadata summaries as distinct if the structures of the sub-trees of the first and second metadata summaries are not equivalent; and

 - if the structures of the sub-trees of the first and second metadata summaries are equivalent, further compare the first and second metadata summaries,

 - wherein the further comparison of the first and second metadata summaries includes:

 - comparing the first and second metadata summaries on a textual level by comparing textual content from the first document that is contained in the sub-trees of the first metadata summary with textual content from the second document that is contained in the sub-trees of the second metadata summary; and

 - identifying the first and second documents corresponding to the first and second metadata summaries as distinct if the textual

content within the sub-trees of the first and second metadata summaries are not equivalent.

14. The system of claim 12, wherein the further comparison of the first and second metadata summaries further includes the:

before comparing the first and second metadata summaries on a textual level, comparing the first and second metadata summaries on an attribute level by comparing attribute values within the sub-trees of the first metadata summary with attribute values within the sub-trees of the second metadata summary; and
identifying the first and second documents corresponding to the first and second metadata summaries as distinct if the attribute values within the sub-trees of the first and second metadata summaries are not equivalent.

17. A program product for use in a computer system that executes program steps recorded in a computer-readable media to perform a method for classifying electronically posted documents, the program product comprising:

a record-able media;
a program of computer-readable instructions executable by the computer system to perform processes comprising the steps of:
receiving a first document and a second document;
generating a first metadata summary for said first document and a second metadata summary for the second document, wherein the first metadata summary includes a first plurality of sub-trees and the second metadata summary includes a second plurality of sub-trees, and wherein each of the sub-trees includes a plurality of nodes;
comparing the first and second metadata summaries on a structural level by comparing a structure of the sub-trees of the first metadata summary with a structure of the sub-trees of the second metadata summary;

identifying the first and second documents as distinct if the structures of the sub-trees of the first and second metadata summaries are not equivalent;

if the structures of the sub-trees of the first and second metadata summaries are equivalent, performing a further comparison of the first and second metadata summaries,

wherein the further comparison of the first and second metadata summaries includes the sub-steps of:

comparing the first and second metadata summaries on a textual level by comparing textual content from the first document that is contained in the sub-trees of the first metadata summary with textual content from the second document that is contained in the sub-trees of the second metadata summary; and

identifying the first and second documents as distinct if the textual content within the sub-trees of the first and second metadata summaries are not equivalent.

18. The program product of claim 17, wherein the further comparison of the first and second metadata summaries further includes the sub-steps of:

before comparing the first and second metadata summaries on a textual level, comparing the first and second metadata summaries on an attribute level by comparing attribute values within the sub-trees of the first metadata summary with attribute values within the sub-trees of the second metadata summary; and

identifying the first and second documents as distinct if the attribute values within the sub-trees of the first and second metadata summaries are not equivalent.

20. The program product of claim 17, further comprising the step of identifying the first and second documents as duplicates if the textual content within sub-trees of the first and second metadata summaries are equivalent.

21. The program product of claim 20, further comprising the step of removing the second metadata summary if the first and second documents are identified as duplicates.

25. A program product for use in a computer system that executes program steps recorded in a computer-readable media to perform a method for classifying electronically posted documents, the program product comprising:

- a record-able media;

- a program of computer-readable instructions executable by the computer system to perform method steps comprising:

- receiving a plurality of documents;

- generating a metadata summary for each of the received documents, wherein each of the metadata summaries includes a plurality of sub-trees and each of the sub-trees includes a plurality of nodes;

- grouping a first subset of the metadata summaries into a first summary group, the first summary group consisting of all of the metadata summaries having a first mime-type designation;

- selecting a first metadata summary and a second metadata summary from the first summary group

- comparing the first and second metadata summaries on a structural level by comparing a structure of the sub-trees of the first metadata summary with a structure of the sub-trees of the second metadata summary; and

- identifying the first and second documents as distinct if the structures of the sub-trees of the first and second metadata summaries are not equivalent.

26. The program product of claim 25, wherein the step of grouping further comprises grouping a second subset of the metadata summaries into a second summary group, the second summary group consisting of all of the metadata summaries having a second mime-type designation.

IX. EVIDENCE APPENDIX

NONE

X. RELATED PROCEEDINGS APPENDIX

NONE